9 Auditory-Visual Speech Processing
Something Doesn't Add Up

ERIC VATIKIOTIS-BATESON AND
KEVIN G. MUNHALL

9.1 Introduction

The multimodal production and multisensory perception of speech have received much research attention in the past 60 years since Sumby and Pollack's landmark demonstration that being able to see a talker's face in noisy acoustic conditions dramatically improves speech intelligibility (Sumby and Pollack 1954). Myriad studies have pursued various conceptual lines about the production and processing of multisensory information in the context of diverse tasks applied to clinical populations and hordes of undergraduate psychology students, and in technical applications for multimedia and speech technology. Previously, we have reviewed the progress in auditory-visual speech processing, particularly with respect to the production and perception of time-varying speech behavior (Munhall and Vatikiotis-Bateson 1998, 2004; Vatikiotis-Bateson and Munhall 2012a, 2012b). In this chapter,[1] we examine what has been learned about auditory-visual speech processing (AVSP) from the potentially disturbing perspective that we still do not have a cogent story for how the visual enhancement of speech intelligibility works. Examining the neural underpinnings of AVSP is, of course, a promising and increasingly well-worn path toward working out a suitable story. However, before turning to neurophysiology to account for behavior, we think it worthwhile to critically review what we have and have not learned from behavioral studies of the production and perception of multimodal speech. In particular, the research questions that have been asked were based on premises and assumptions about language and cognition that may need to be rethought before the observed results can begin to make sense, and should be reexamined

The Handbook of Speech Production, First Edition. Edited by Melissa A. Redford.

and possibly reframed or discarded entirely before looking for answers in the neural processes of the brain. In the following, we first summarize the findings about which we are confident. Then, we discuss findings that either lead to conflicting interpretations and/or cast doubt on the meta-theoretical premises and assumptions that shaped the way the research questions were formulated. We conclude the chapter with a tentative prescription for future studies based on new questions about AVSP.

9.2 What we think we know and think we understand

In our own studies of the production and perception of multimodal speech, carried out over the past 20 years, we have discovered or confirmed a number of facts about which there appears to be little controversy insofar as the results have proven to be robustly replicable across myriad differences in experimental methodology and design. How they all fit together and should be interpreted is a more difficult problem and is taken up in sections 9.3 and 9.4.

9.2.1 Causal and functional linkages in multimodal speech production

Configuring the vocal tract through time simultaneously shapes the acoustic resonances of the speech signal and visibly deforms the face, primarily through the motions of the jaw and shaping of the lips. That is, what happens in the vocal tract during speech production *physically* determines the audible and visible signals that result from that process. Simply recognizing that the face *defines* the sidewalls of the vocal tract should lead one to expect a tight, causal coupling between the vocal tract and the face. We demonstrated this coupling with analyses of kinematic (vocal tract and face) and acoustic data for speakers of Japanese and English published in the mid-to-late 1990s with Takaaki Kuratate, Philip Rubin, and Hani Yehia (Vatikiotis-Bateson and Yehia 1996b; Yehia, Kuratate, and Vatikiotis-Bateson 1999; Yehia, Rubin, and Vatikiotis-Bateson 1998).

In these studies, analyses of Japanese and English sentence production showed that measures of two-dimensional (2D: midsagittal height and protrusion) position of the jaw, lips, and four flesh points on the anterior tongue surface correspond closely with measures from the three-dimensional (3D) positions of markers – usually 17 or 18 – arranged on the chin, lips, and other regions of the face below the eyes (see Figure 9.1). Calculated as a multivariate correlation averaged over the span of medium-length sentences (20+ syllables), the correspondence between the measures made in the vocal tract and on the face was strong enough to estimate more than 80% of the face motion behavior from vocal-tract articulation. Similarly, but not quite as efficiently, about 65% of the spectral acoustics could be estimated from vocal-tract articulation, with the

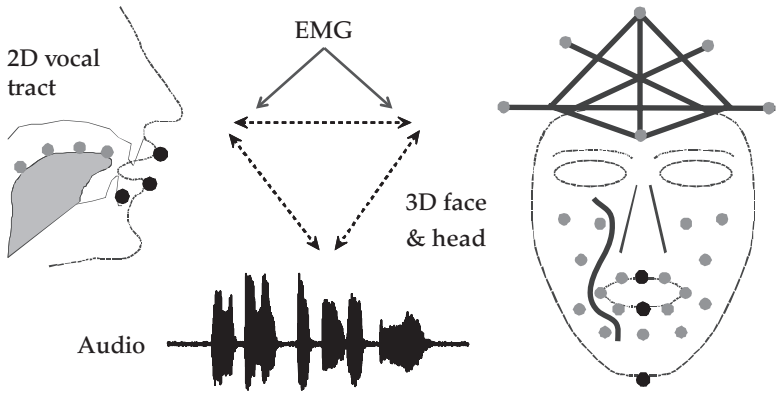

Figure 9.1 Schematic of physiological and acoustic production measures: muscle activity (EMG), 2D vocal-tract motion, 3D face motion, 6D head motion, and the audio acoustics. The arrows indicate the various cross-domain analyses that have been carried out.

frequencies in the vicinity of the second formant band (F2: 1500–2500 Hz) estimated at 80% accuracy or better. This last finding is not surprising since F2 corresponds most closely with the shape of the front oral cavity, which is where the electromagnetic tongue markers were located (for details about the EMMA system, see Perkell et al. 1992).

Applying the elementary principle that two things that are each similar to a third should be similar to each other (i.e., if $a \approx c$ and $b \approx c$, then $a \approx b$), these analyses also showed that 3D face motion could be estimated from the spectral components of the acoustics at about 95% accuracy or better using nonlinear estimation techniques (Yehia et al. 1999; Yehia, Kuratate, and Vatikiotis-Bateson 2002). In these analyses, the correspondence between face motion and acoustics was given a significant boost by small deformations in the face that were likely due to structured fluctuations of intraoral air pressure (see Carter, Shadle, and Davies 1996).[2] Although small in terms of the effect on signal variance, these structural fluctuations have proven highly effective in subsequent synthesis of speech based on nonlinear estimates of time-varying acoustic spectral parameters from facial markers located bilaterally on the cheeks (see Figure 9.1).

In a subsequent study of Japanese and English sentence production, in which no movement-restricting vocal tract measures were made, a *functional* linkage was identified between the rigid body (6D) motion of the head and the fundamental frequency (F0) of the speech acoustics (Yehia et al. 2002).[3] Correlation analysis showed that, as F0 increases, the head tends to tilt upward and away from the chest; and downward (chin closer to the chest) as F0 decreases. That the linkage can be decoupled with practice supports the notion that this is a functional linkage rather than something primarily structural. For example, while most people have

difficulty damping or reversing the relation – try it for yourself – trained singers have no problem decoupling head motion and F0.

Another functional correspondence between head motion and the speech acoustics is the tendency for head motion to increase with acoustic amplitude (intensity). In early motion studies where head motion was an undesirable artifact, the head was usually constrained by a headband of some sort. This had the unpleasant effect of reducing the talker's vocal amplitude. As reported by Tom Scholte (Department of Film and Theatre, University of British Columbia) and confirmed repeatedly by others involved in theatre, decoupling head motion and vocal amplitude is a basic component in training actors to project their voices at higher amplitudes while maintaining the visual demeanor appropriate to the much lower amplitude appropriate to, say, face-to-face interaction. On the other hand, the repeated observation that acoustic amplitude (root mean square: RMS) also correlates with movement amplitude of vocal tract and facial motion suggests a physical, rather than functional, coupling (Barbosa et al. 2006; Barbosa, Yehia, and Vatikiotis-Bateson 2008; Yehia et al. 2002).

9.2.2 Visible actions relevant to speech intelligibility are everywhere

In one of our early forays into AVSP, we conducted a study whose aim was to determine where Japanese or English listeners direct their eyes (foveate) during audiovisual perception and the extent to which their eye motion patterns change under different auditory and visual conditions (Vatikiotis-Bateson et al. 1998). We recorded the eye movement behavior of perceivers watching video monologues presented with a range of amplitudes of auditory masking noise (a multilingual party recorded in a Japanese kitchen) and at image sizes ranging from normal (for face-face interaction at approximately 1 meter inter-talker distance) to much larger than normal. Simple questions at the end of each trial such as "Did he say '*peep*' or '*beep*'?" served to index the stimulus intelligibility while focusing perceiver attention on phonetic aspects of the talker's speech rather than on other factors such as the talker's sincerity or happiness (Eigsti et al. 1995). The basic findings of the Vatikiotis-Bateson et al. (1998) study were that, even under conditions of the highest masking noise and largest image projection, perceivers foveate more on the eyes and less on the mouth than previously believed. Also, eye movement patterns – specifically saccades between the eyes and the mouth – change little, if at all, even when the noise levels are high and the image size so large that the eyes and mouth cannot both be viewed within the relatively high-acuity region of the perifovea.

It was clear from pretesting, aimed at determining the appropriate noise levels for auditory-only conditions (Vatikiotis-Bateson, Eigsti, and Yano 1994), that being able to see the face enhances intelligibility. This was expected from Sumby and Pollack's (1954) earlier research, at least when the faces are displayed at sizes appropriate for face-to-face interaction. What was surprising was that the same

enhancement was observed even at the largest image sizes, when perceivers could not simultaneously foveate on the eyes and keep the mouth in sharp focus. This led to two questions that were pursued in subsequent production and perception studies:

- How much spatial and temporal acuity (or resolution) are needed for visual enhancement of speech intelligibility to occur?
- Where is the linguistically relevant information located?

Based on what we thought we knew in the 1990s about the role of eye motion in optimizing visual acuity – namely, that spatial acuity was highest at the fovea and temporal acuity was highest at the visual periphery (Carpenter 1988) – our results for perceiver eye motion during distorted visual and noise-degraded auditory speech perception suggested two alternative hypotheses about the location of visual information on the face. First was the hypothesis that linguistically relevant visual speech information is distributed widely across the face, rather than merely in the vicinity of the mouth. Earlier analysis of orofacial production data supported this hypothesis (Vatikiotis-Bateson and Yehia 1996b). For semi-spontaneous production of short sentences and phrases, the correspondence between two sets of 3D facial motion markers was extremely high (see Figure 9.1 for delineation of the two marker sets). The inner set consisted of five markers placed around the right-hand perimeter of the lips (sagittal midline of the upper and lower lips, the right-hand corner of the lips, and two markers midway between the midline and corner on each lip). The outer consisted of five markers away from the mouth on the right side of the face (two on the lower face and three on the cheeks). Marker motion for the inner set was highly predictable, $85\% < r^2 < 99\%$, from motion of the outer set.

The alternative hypothesis, based on the acclaimed temporal acuity of visual periphery, was that temporal information is more important than spatial information for the visual enhancement of speech intelligibility. This made sense, given perceivers' strong performance recovering speech information presented at high levels of auditory masking noise and at large image sizes, where foveating on the stimulus talker's eyes put the mouth region 10–11 degrees away from the foveal center.

As it turned out, neither hypothesis about how perceivers make use of temporal and spatial information in visible speech was borne out by subsequent studies.

9.2.3 Very little visual information is required for perception

In a series of SPeech-In-Noise (SPIN) studies, Munhall and colleagues showed that perceivers retrieve relevant visual speech information at low spatial and temporal resolutions (dePaula et al. 2003, 2006; Munhall, Jozan, et al. 2004; for overview, see Munhall and Vatikiotis-Bateson 2004). These studies showed that perceivers could retrieve visual speech information presented in noise at cumulative (lowpass

filtered) spatial frequencies as low as 5 cycles per face (cpf), using a Chebchev filter (dePaula et al. 2006), and at 7 cpf, using one-octave passband filtered images for a range of center resolutions of 3–44 cps (Munhall, Jozan, et al. 2004).

To determine the lower bound on temporal resolution of the visual information, dePaula et al. (2006) used Gaussian filtering of the image sequence to show that intelligibility of semantically unpredictable sentences (based on specialized topics and vocabulary) did not begin to degrade until temporal resolution fell to 6–9 Hz (depending on listener). These values are substantially lower than the 14–16 Hz previously reported by Vitkovich and Barber (1994). We attribute the discrepancy to the fact that Vitkovich and Barber used frame decimation which, by removing frames from the image sequence, reduces the frame rate and increases the duration of the black gaps between frames. These gaps, as they get larger, could disrupt processing of the visual information. The Gaussian filter used by dePaula et al. (2006), on the other hand, reduced the temporal resolution without reducing the frame rate, which remained at 30 fps. One drawback to the Gaussian filter method is that smearing the reduced temporal information across 30 fps also reduces the spatial information represented in each frame, making it difficult to determine the exact contribution of temporal resolution to perceptual performance.

Finally, in their study of spatial resolution requirements for visual enhancement of speech intelligibility, Munhall, Jozan, et al. (2004) also showed that the effectiveness of visual speech information in enhancing the intelligibility of speech produced with noise-masked acoustics is not affected by the relative size of the talker's image on the perceiver's retina. They rigorously tested talker–perceiver distances between 1 m and more than 3 m and found no degradation in perception. At the time, Munhall, Jozan, et al. inferred from this finding that the visual enhancement must be cognitive rather than physical.

9.3 What we know, but do not understand

In this section we discuss different pieces of research whose results are solid enough, but which have not yet lent themselves to tidy interpretation, especially when it comes to connecting the results of production studies with multisensory perception studies. For example, as discussed in section 9.2.1, the motion of the head correlates well with F0 and to a lesser extent with acoustic amplitude. One would hope that perceivers take advantage of such strong production linkages, and it appears that they do; but, as we discuss in section 9.3.1 below, more questions are raised than answered by the finding that being able to see the head's motion during speech contributes substantially to the perceived intelligibility of audiovisual stimuli consisting of talking head animations and noise-degraded acoustics.

Even more unsettling, we have not been able to determine crucial orofacial landmarks for measuring facial motion relevant to audiovisual production-perception. Discovering where the relevant information is on the face was one of the original reasons for launching this entire research paradigm in the early 1990s. One motivation for examining perceiver eye movement during multisensory speech

production was the hope that where perceivers foveate would help us identify ideal locations for marker placement. Subsequently, we manipulated marker placements and dimensionality and found that various locations, such as the cheeks and non-midsagittal placements on the lips, made specific contributions to the correspondence with the spectral acoustics (from movement of the cheeks) and the position of the tongue (from movement of the non-midsagittal lips); but we found nothing that clearly contributed to audiovisual speech perception. To make matters worse, it turns out that all of the motion in a large region of interest (ROI) such as the head and face can be reduced to one time-varying magnitude with little loss of relevant information about the spatiotemporal organization of the speech behavior. We discuss this latter issue further in section 9.3.2.

Finally, the idea that speech production is a nonlinear, distributed process is not new. One of the most successful attempts to link speech performance and linguistic structure, Articulatory Phonology (for early overview, see Browman and Goldstein 1986), construes speech production as a symphony of semi-independent articulatory events, temporally orchestrated to attain serialized linguistic goals. Indeed, were it not for the predominant conceptual dependence on the notion that the continuous speech stream must be decomposable into strings of contrastive phoneme segments, the fact that information crucial to speech perception is distributed over substantial spans of signal would not be surprising at all. In section 9.3.3, we discuss one aspect of the distributed timing of events specific to the production and (non)perception of the labial viseme; in part because it further emphasizes the disconnect between information contained in speech signals and how perceivers access that information, and partly because we want to emphasize that human perception performance should not be used as a gold standard for building artificial perception systems.

9.3.1 Head motion provides crucial information for audiovisual perception

It is well known that the head is an active and important component of communicative interaction, providing paralinguistic information (see Trager 1958) including emphasis and indications of listener attention, comprehension, disagreement, and the like. The finding that head motion correlates well with F0 suggests that the head potentially provides perceivers redundant information that might otherwise be lost when the acoustic signal is severely degraded. This is important because F0 may contribute segmental information about vowel identity – insofar as different sonorants have different intrinsic F0 ranges (e.g., Ewan 1979; Vilkman et al. 1989; Katz and Assmann 2001) – and conveys substantial prosodic information about stress and intonation via its modulation over the course of an utterance.

We say the head *potentially* provides information to perceivers because the existence of such redundancies or any of the other measured correlations in the production data is no guarantee that perceivers actually detect and make use of them. This is why we went to so much effort to create a talking head animation system that could be used to create synthesized stimuli from time-varying physiological and acoustic data for perceptual evaluation (Kuratate, Yehia, and

Vatikiotis-Bateson 1998). In a study designed both to determine the linguistic validity of the animated stimuli for perceivers and to examine the extent to which head motion influences audiovisual speech perception, Munhall, Jones, and colleagues (2004) presented talking head animations in auditory noise-masking for three conditions: the motion of the face and the head synthesized from recorded kinematics of these structures, the motion of the face without head motion, and motion of the face and head with doubled amplitudes, but no change in movement times (thereby doubling the velocity of all motions.)[4]

The intelligibility results for Japanese semantically unpredictable sentences, presented with noise-masked acoustics, are shown in Figure 9.2. As shown, intelligibility was reliably better for all three talking head conditions than for noisy audio alone. The best results were obtained for the head + face condition, which was the most natural of the video conditions. Distorting the spatiotemporal acoustics disrupted audiovisual intelligibility the most. The results for normal face motion without head motion were closer to the results for the distorted kinematics than for the combination of face, head, and noisy audio.

The talking head animation system provided a first demonstration that seeing rigid body motion of the head enhances speech intelligibility substantially. It simply is not possible to produce natural, communicative speech without head motion, and certainly the face motion alone condition is not natural insofar as no typical speaker holds the head completely still while producing speech at normal

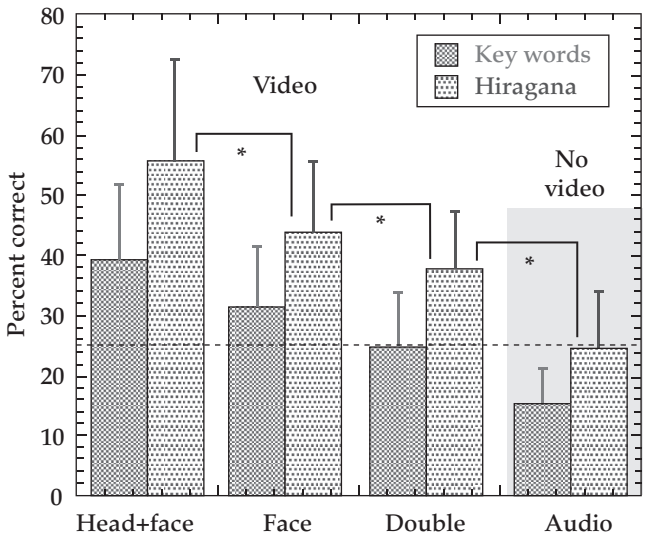

Figure 9.2 Intelligibility of semantically unpredictable sentences (Japanese) presented in acoustic masking noise for one, noise-masked audio control and three audiovisual conditions (head + face, double, face presented with noise-masked audio). Results are shown for the percent correct of key words and syllables (hiragana). Modified from Munhall, Jones, et al. (2004: Figure 3).

levels. As for why head motion enhances intelligibility, this study does not help provide a real answer. With the exception mentioned before of some possible vowel quality information gleaned from inherent F0, there is surely no substantial segmental information accessible by monitoring head motion. Instead, it is more likely that head motion plays one or more roles in conveying speech prosody. Modulation of head motion in rhythmically tuned chunks, such as the stress foot in English, prominence peaks, etc., could help perceivers align to the multisensory speech signal, which in turn could enhance auditory and/or visual perception. That is, the head's role in audiovisual perception could be indirectly a matter of helping listeners find the signal(s) from which they need to glean speech information. A visual analog to this is watching telephone poles from the side window of a fast-moving car or train. If you track the motion of the poles with your head and eyes, you obtain much more information about the poles than when you stare straight out the window and let them whiz across your retina.

In sum, while there is no question that head motion enhances audiovisual speech intelligibility, we do not have much idea about how this enhancement is achieved. We suspect that the visible head motion alone may contribute to entraining listeners to the modulated acoustic signal, but this is not sufficient to induce observable enhancement of speech intelligibility (suggested by the failure of animations containing only head motion and degraded acoustics to improve intelligibility: Hill and Vatikiotis-Bateson 2005). More likely, listeners require the combined motion of the head and face *and* their knowledge of the specific language prosody to align to the signal. We return to the question of perceiver alignment to production in section 9.3.3.

9.3.2 Crucial orofacial landmarks are difficult to determine (almost anything works)

When Yehia and colleagues originally calculated the multivariate correspondences within and across measurement domains (Vatikiotis-Bateson and Yehia 1996a; Yehia et al. 1998; Yehia et al. 1999, 2002), their concern was first to get high-resolution spatiotemporal measures for as many physiological channels as possible, especially for the face and head. Once obtained, they used filtering and dimensionality reduction techniques, such as principal component analysis (PCA), to optimize the complexity of the data. This approach made sense at the time, because speech-related signals had never been examined this way before. There was also a craze, encouraged by new technology for acquiring high-resolution signals in abundance, for oversampling time-varying spatiotemporal data. In the end, this work established that cross-domain estimations between vocal-tract articulation, face and head motion, and acoustics are temporally fitted to the rate of opening and closing the vocal tract (under 9 Hz), and that the within-domain spatial complexity can be reduced to five or six orthogonally independent components for vocal tract and orofacial motion and 10 components for the spectral acoustics (plus two more for F0 and amplitude).

The reduction in complexity to five or six components for vocal-tract and orofacial motion is consistent with the finding discussed in section 9.2.3 above that perceivers require very little detail in the visual stream to enhance utterance intelligibility. The low dimensionality of visual speech information coupled with the finding that linguistically relevant information appears to be distributed all over the face led Barbosa and colleagues to conduct a more careful consideration of what needs to be measured and where. In a series of studies beginning with his PhD thesis, Barbosa showed that 2D measures were just as good as 3D measures for within- and cross-domain analysis (Barbosa 2004). This finding facilitated the transition from tethered marker or difficult-to-calibrate passive marker systems to physically non-invasive, video-based recovery of visible 2D motion (Barbosa and Vatikiotis-Bateson 2006; Barbosa et al. 2006; Barbosa, Yehia, and Vatikiotis-Bateson 2003). These studies varied the number, placement, and physical characteristics (e.g., size, color, whether painted or pasted on) of markers and tested different algorithms for marker tracking.

Finally, we compared markerless motion measures derived from video using Horn and Schunk's (1981) optical flow algorithm with those obtained using marker-based systems. Although it seemed absurd at first, test after test showed that most analyses spanning multiple measurement domains, in which the correspondence between domains is the primary interest, can be carried out with little or no change in outcome using only one ROI for the entire head and face. Specifically, there is almost no difference in estimation power between measuring 25 individual markers on the face and head, and summing the changes of pixel intensity, converted by optical flow analysis to changes of position (i.e., velocities) between one image frame and the next, into one time-series of all the motion within an ROI (Barbosa et al. 2008). Even though many analyses are improved by keeping the horizontal (x) and vertical (y) components of the optical flow distinct, analyses focused on computing continuous correspondence with other measured behaviors do equally well or better by conflating the x and y components to one, time-varying, Euclidean amplitude (Barbosa et al. 2012).

By way of contrast, consider the work done in the 1970s and 1980s by Paul Ekman and others leading to the Facial Action Coding System (FACS: Ekman and Friesen 1978). The aim of that work was not to model an individual's time-varying behavior across a range of speaking contexts via analysis of image sequences; rather the aim was to establish consistent locations and number of landmarks across sets of isolated images for different people expressing specially concocted emotions such as happiness, sadness, and anger (Ekman 1989; Ekman and Friesen 1978; Ekman, Friesen, and Ellsworth 1972). In the 1990s, the Institute of Electrical and Electronics Engineers (IEEE), and other international bodies concerned with developing standards for digital video, finally settled on a set of nearly 70 facial landmarks taken from FACS to incorporate into MPEG-4, currently the prevalent video compression format used in science, technology, and commercial applications, including mesh-based animations (e.g., Tekalp and Ostermann 2000). Although industrial MPEG-4 animations tend not to receive rigorous perceptual evaluations of the sort applied to the talking head animation system developed by

Kuratate and colleagues (2005, 1998), the low dimensionality and the demonstrated low spatial and temporal resolution required for most multimodal analyses and particularly multisensory perception suggest that the FACS codes of MPEG-4 are more than sufficient, possibly even excessive, for most research applications.

9.3.3 The temporal distribution of linguistically relevant audiovisual information

Almost anyone who has studied speech phenomena that straddle the boundary between phonetics and phonology has heard of anticipatory and carryover coarticulation in which some attribute such as nasalization of stop consonants, /m, n, ŋ/, or rounding of high back vowels, /u, o/ in English, is observable for some amount of time before and/or after the presumed production of the associated phoneme. That is, whether the velar port, connecting the oral and nasal cavities is open or closed does not interfere with the identification of other phonemes that have no nasal counterpart (e.g., Bell-Berti et al. 1979). Similarly, lip rounding, which has been treated as a secondary feature for English vowels classified as high and back, does not interfere with the production of most other consonants and vowels (e.g., Bell-Berti and Harris 1982). In both of these examples, the audible and visible (if there are any) components of the production co-occur continuously before, during, and possibly after their supposed phonemic moment.

A quite different example of temporal distribution involves a temporal dislocation of the acoustic consequences from the articulation that produces them; that is, in many contexts and languages, plosive stops such as /p, t, k, b, d, g/ have no acoustic realization during the articulation of the consonant and are audible only after stop release during the transition to the onset of the vowel and possibly during the transition from a preceding vowel into the stop closure (Catford 1977). From the perspective of audiovisual production and perception, this becomes quite interesting because the temporal dislocation between the visible articulation of a bilabial plosive, /p, b/, and its auditory consequence is roughly 150–200 ms (Abry, Lallouache, and Cathiard 1996; Cathiard, Lallouache, and Abry 1996). In other words, perceivers must align with multimodal signal components displaced substantially in time.

Munhall and colleagues (Munhall et al. 1996; Munhall and Tohkura 1998) examined a version of the temporal alignment problem using McGurk effect stimuli (McGurk and MacDonald 1976). The McGurk effect, also known as the "fusion illusion," pertains to the mandatory integration of mismatched visual and auditory speech stimuli, resulting in a percept different from either of the original stimuli. Munhall and colleagues showed that perceivers could fuse auditory /ba/ and visual /ga/ stimuli to perceive something like [da] across substantial temporal dislocations. Interestingly, the tolerance for temporal dislocation was much greater when the audible component followed the visible component. This asymmetry accommodates the increased temporal displacement that occurs when perceivers are further away from a sound source. It also fits the causal sequence of a physical

(e.g., vocal tract) event having acoustic consequences, rather than the other way around: just as we do not expect thunder to precede lightning, we do not expect sound to precede movement.

Another way to consider the temporal distribution of audiovisual events is that the distributed presentation of related and/or redundant events facilitates perception by providing perceivers a better opportunity to align to the multisensory stream. This is what we proposed in section 9.3.1 for the prosodic role of head motion. Because related and/or redundant event streams are primarily a consequence of how the physical system is organized and behaves, there is no guarantee that perceivers will, in fact, develop or commit the cognitive resources needed to take full advantage of the opportunity these time-varying events provide. As a preliminary test of the potential mismatch between production events and perception, several students designed an audiovisual production and perception study to test the extent to which /p, b, m/ can be distinguished via computational analysis of the visible face and head motion, even if not perceptually (Abel et al. 2011). The three bilabial stops form the classic labial viseme in which the component sounds are not visually distinguishable (Woodward and Barber 1960). In earlier work, the high confusability of these stops was tested on static key frames, rather than on image sequences (for overview, see Bruce and Young 1986); but the student study showed that viseme-internal differences cannot be perceived even when presented in short image sequences excised either from nonsense VCV sequences (e.g., aba, ama, apa) or from different positions in sentential utterances such as, "it was the sabby/sammy/sappy that went to the store." This is not to say that perceivers are entirely insensitive to differences between the bilabial stops. Most subjects in the student study had more difficulty correctly identifying the visibly and audibly most neutral /b/ productions than one or both of the other two labials. This response bias suggests a kind of viseme-internal discrimination that is simply too weak to reliably differentiate the bilabial stops.

The story is quite different when the production data are considered. Applying optical flow analysis (Barbosa et al. 2008; Horn and Schunk 1981) to a single ROI that encompassed the entire head and face showed reliable differences in the time course and amplitude of motion within the ROI associated with the three labial stops. Similar to the temporal distribution of nasality discussed previously, stop-specific differences in visible motion spanned substantial stretches of signal that included, for example, the transitions from the vowel preceding the word, sabby, into the initial /s/, from the /s/ to /ae/, from the /ae/ to /b/, and from /b/ to /i/. These differences were reliable for both talkers and confirmed the much earlier observation that the voicing of a final obstruent in simple CVCs such as [baeb] and [baep] influences the kinematics of both the CV and VC transitions (Kelso et al. 1984; Vatikiotis-Bateson and Kelso 1984).

In sum, this study provides a clear example of perceivers not being able to exploit differences contained within the production data. That is, the perception results confirmed the visual confusability expected of the three stops comprising the labial viseme, even though the head and face motion contained measurable, temporally distributed differences associated with the three labial stops. On a

happier note, these results demonstrate that, despite the inability of human perceivers to distinguish the visible differences between labial stops, a machine recognition system would have no trouble.

9.4 Recommendations for future studies

In the 20 years that we have been involved with audiovisual speech research, we have had some remarkable success that just scratches the surface of our original question about what the visual channel contributes to speech perception. Our results certainly have raised many more questions than we are likely to answer in what remains of our careers. Rather than try to enumerate these as a shopping list, we describe several broad lines of inquiry that we believe could greatly increase our understanding of auditory-visual speech processing. Rigorous future research is needed to

- attain a better understanding of the role of redundancy in both the production and perception of speech,
- take a much closer look at brain function during audiovisual speech perception,
- assess the role of spatiotemporal coordination in the production and perception of audiovisual speech.

9.4.1 Redundancy in AVSP

Redundancy is essential to the successful operation of many systems, both natural and artificial. Vertebrates are equipped bilaterally with pairs of limbs, sensory organs, two brain hemispheres, along with pairs of some other internal organs (lungs, kidneys, etc.). Artificial systems, in which the consequences of system failure are deemed unacceptable, such as the flight controls that keep airplanes in the air, are replicated – many times in the case of commercial aircraft – with independent systems that can be called into service when a primary system fails. Yet, in science, systems are modeled by specifying the smallest number of parameters needed to characterize or simulate the system's structure and/or observed behavior. What gives? Why is it that in many branches of scientific inquiry, optimization is synonymous with maximum parsimony, and the modeling process incorporates greater complexity only when it is demonstrably necessary?

Language research has suffered from this minimalist approach to optimization. For example, phonemics, whose business in the first half of the twentieth century was to describe and classify the sounds of a language, relied heavily on establishing the set of linguistically contrastive elements, or phonemes, whose descriptions included two types of feature sets: *distinctive* features which contributed to distinguishing one phoneme from all others, and *descriptive* features which collated all of a phoneme's known attributes (Jakobson, Fant, and Halle 1951/1963). By the time Chomsky and Halle's *Sound Pattern of English* was published (1968), descriptive

features were beginning to be called redundant features with the negative implication that they were not needed for linguistic analysis, and certainly not for specifying the paradigmatic contrasts within a particular phoneme inventory (for the cybernetic/information theoretic precursor to this formalization, see Ashby 1956). In contrast, a vast amount of speech research has focused on discovering the why and how of sounds interacting syntagmatically – in the speech stream as it unfolds in time. Indeed, without recognizing it as such, much of the research on speech synthesis and recognition, where coarticulation and other processes such as dissimilation span multiple segments, has had to depend at least as much on the descriptive attributes of sounds as on their contrastive features.

In audiovisual speech research, determining what is redundant in the acoustic and visible signal streams has never been systematically investigated. In large part, this has to do with the difference of perspective distinguishing production and perception research. As one would expect, most research proceeds from an observation of behavior to questions about how the behavior came about and what effects it has on other behavior. In other words, the initial perspective is perceptual. In spoken language, there is another important bias, namely, that speech is primarily an acoustic/auditory phenomenon. There is no question that whatever contribution is made by the visual channel, it is secondary to that of the auditory stream. Early recognition that there might be useful visual information most certainly arose in contexts of hearing deficits and situations where environmental noise masked speech acoustics. It is then not surprising that early audiovisual research by Quentin Summerfield treated the visual stream as a source of information that complemented the more fragile aspects of the acoustics, such as rapid transitions in higher frequency acoustics and their relation to visible changes in vocal tract configuration. Initially, Summerfield (1979) proposed that the complementarity of different sensory channels might be orchestrated cognitively, but he backed away from this view later (Summerfield 1987). The definition and status of complementarity in audiovisual perception has evolved along lines more amenable to the research described in this chapter (e.g., Grant and Seitz 2000).

The work by Yehia and colleagues, discussed earlier, exploring the correspondences between the various audible and visible signal domains was by definition largely dependent on redundancy between the domains. That the motion of the head and face can be used to synthesize more or less intelligible speech acoustics, and the acoustics can be used to synthesize very accurate head and face motion, leaves little doubt about the importance of redundancy in audiovisual production. Furthermore, we know that much of the audiovisual correspondence stems from a common source – vocal tract articulation. The other major sources of redundancy are due to the functional coupling of the head to vocalization, which is itself not entirely decoupled from the vocal tract dynamics despite the mathematically convenient fantasy about the independence of the vocal *source* at the larynx and the vocal tract *filter* (Fant 1960).

Indeed, in speech production, complementarity implies *uncorrelated* signal components, but this simply has not been examined. Similarly, in speech perception, complementarity arising from redundant multimodal production of the sort discussed

here needs to be distinguished from complementarity created by selectively attending to and combining otherwise unrelated signal events. Doing so might, for example, better inform us about the processes underlying various phonetic and non-phonetic forms of convergence and divergence that occur as a result of speaker interaction (e.g., Kim, Horton, and Bradlow 2011; Pardo 2006).

9.4.2 Brain function and AVSP

In the late 1990s, we began to look more closely at the perception side of auditory-visual speech processing. Callan and colleagues (2002, 2004) designed one of the first series of studies examining brain behavior during audiovisual speech perception. Using electroencephalography (EEG) and functional magnetic resonance imaging (fMRI), these studies established a linkage between motor activation and multisensory perception (Callan et al. 2002) and delivered early evidence that visible speech activates auditory areas even when presented without an acoustic signal, while the reverse – activation of visual signals from only auditory stimulation – does not occur (Callan et al. 2004).

A possible interpretation of these findings harkens back to our earlier discussion of the role of visual speech information in speech perception, the notion that the auditory stream is primary, and to the body of evidence our extended group has collected showing how minimal the requirements are for visual enhancement of speech intelligibility to occur. Specifically, the primary function of visual speech information may be to increase the sensitivity of the auditory system. Perhaps coincidentally, it is commonly observed that, when a speaker suddenly becomes visible during a noisy acoustic presentation, the speech signal is perceived to suddenly become louder and clearer. Whether or not such observations are related to somewhat nebulous differences in the activation of the auditory system, the possibility that visual speech information boosts the gain of the auditory system deserves closer attention. Doing so would not only tie together many of the loose ends in our research, as discussed in this chapter, but would also put an entirely different perspective on other longstanding issues, such as the relevance of auditory-visual integration in speech processing (e.g., Massaro and Cohen 1983; Robert-Ribes, Schwartz, and Escudier 1995).

9.4.3 Spatiotemporal coordination in audiovisual speech

An underlying theme of this chapter has been a general concern about how spatial and temporal coordination operates at various levels of observation within and between individuals during communicative interaction. Our discussion in sections 9.3.1 and 9.3.3 about how perceivers must first align with multimodal signals in order to actually retrieve relevant speech information is but one piece of this concern. Another issue is how perceivers, who are also typically producers of the languages they perceive, manage the re-characterization of event timing at different levels of observation. For example, if we accept for a moment the multi-tiered orchestration of articulatory events proposed by Articulatory Phonology, the

timing of gestural scores is clearly quite different from the temporal scheme of the subsequent acoustic signal. Aided by literacy (and perhaps some linguistic training), we readily discern a much more linear stream of events, consisting of modulated rises and falls in acoustic amplitude coinciding generally with opening and closing the vocal tract, which in turn signal syllable nuclei, different manners of consonant production, and the like.

Our own work on multimodal speech production readily coincides with the gestural score approach outlined by Articulatory Phonology. We have demonstrated that the strong correspondences between vocal tract, orofacial, and acoustic signals described in section 9.2.1 capture at least some of what is needed to bridge between production and perception. In particular, we have synthesized fairly complex and appropriate acoustics from the motion signals of the face and head, and have done an even better job in the other direction, from acoustics to visible motion, due to the greater richness of the measured acoustic signal compared to the motion signals being estimated (for discussion, see Yehia et al. 2002). However, these analyses are computational and driven entirely by time-varying signals; they do not involve phonemes or any other construct hypothesized to be associated with linguistically tuned perception. From the computational perspective, this is probably a good thing. Recent advances in machine recognition of continuous speech using deep belief networks (Hinton, Osindero, and Teh 2006) have outperformed phoneme-constrained approaches to automatic speech recognition (ASR) that combine phoneme-aligned Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs).

Despite the success of the various signal-based computational approaches in bridging between measurement domains without having to rely on constructs such as phonemes, we worry that the organization and timing of multimodal events is not only physical, but also cognitive; as is perception. Multisensory perception suggests sensitivity in the visual domain to the multi-tiered timing of the jaw, lips, head, and even changes in intra-oral air pressure. Even though the subsequent acoustic signal is related to articulatory behavior in a straightforward, if somewhat complex way, the acoustics may activate a different temporal structure in the auditory system. If so, how are these different timing regimes managed? We think this will be a difficult question to answer until more is known about the nature of and connection between physical and cognitive constraints on production and perception.

Because biological systems are anything but efficient and parsimoniously given to singular descriptions, it may be that there is no formalizable connection between the physical and cognitive aspects of speech processing. This could partially explain why perception does not always fully exploit the output of production. Unlike the iterative algorithm in a computational analysis that can take as long as it needs to process a speech event, the perceptual system has one shot at a unique signal perceived in a unique context. Cognition is notoriously efficient at making error-prone predictions about novel events. The work of Abel et al. (2011) shows clearly that there are reliable differences in the production of the different English labial stops that are distributed over relatively long temporal spans, but also that perceivers cannot quite grasp these differences, at least not to the point of pushing the correct button key during an experiment. Of course, there may be other

extenuating factors that need to be examined; for example, the tipping point for successful detection may be how familiar the talker is to the perceiver. If the perceiver has prior experience, analogous to the iteratively applied computational algorithm, the perceptual system may do a much better job of aligning to and processing novel instances. Another factor may be that idiosyncratic differences in production are more or less easily parsed by perceivers who are themselves representative of both production and perception asymmetries. It is well known that some talkers produce more readily intelligible speech than others, but we do not yet know what makes one talker more intelligible than another.

9.5 Conclusion

In this chapter we have presented an overview of research on auditory-visual speech processing with which we have been directly involved for the last 20 years. We attempted to identify questions that are either raised by the research and/or must be addressed to clarify the interpretation of results. The findings discussed in section 9.2 are ones about which we are fairly confident because they have proven to be readily replicable and depend little, if at all, on the persistence of any particular theoretical perspective. The studies discussed in section 9.3 are ones that have produced provocative results whose interpretations more actively demand further exploration of questions, such as those raised in section 9.4. More specifically, the past 20 years have seen a general acceptance of the notion that the analysis of spoken communication, like any behavior, must acknowledge that much of its structure is context-specific and variable in time. In this chapter, we push one step further by proposing that production and perception may interface with the same dynamical system, but they do so differently, and that this difference needs to be understood before we can make pronouncements about, for example, the visual contribution to speech intelligibility in noisy acoustics.

NOTES

1 The work reported here was supported primarily by ATR International (Japan) and secondarily by Canadian Tri-Council grants from the Social Sciences and Humanities Research Council (SSHRC) and the National Science and Engineering Research Council (NSERC), the Canada Research Chairs program, and the Canada Foundation for Innovation (CFI). Many students and colleagues have contributed substantially and generously to this research program through the years, too many to list here, but particular thanks goes to the late Yoh'ichi Tohkura, without whose vision and support this research would never have been undertaken.

2 The Optotrak (NDI, Inc.) measures position changes accurately at 0.1 mm, well within the resolution range of the structured light measurement techniques used by Carter et al. (1996) to correlate facial deformation with intraoral air pressure.

3 The three-dimensional motion of rigid objects such as the head or skeletal segments (e.g., arms and legs) consists of six geometric degrees of freedom: three translations defining position and three rotations defining orientation; that is, one translation and one rotation for each of the three Cartesian axes. Unlike deformable objects such as the tongue or lips, the position and orientation of any point on the object has a fixed relation to the position and orientation of every other point on the object; so the entire object can be treated as a single six-dimensional point. For greater detail, see Vatikiotis-Bateson and Ostry (1995).
4 Velocity = distance / time. Therefore, doubling the motion amplitude for the same time period, doubles the speed (velocity) of motion as well.

REFERENCES

Abel, Jennifer, Adriano V. Barbosa, Alexis Black, Connor Mayer, and Eric Vatikiotis-Bateson. 2011. The labial viseme reconsidered: Evidence from production and perception. In Y. Laprie and I. Steiner (eds.), *9th International Seminar on Speech Production (ISSP),* 337–344. CD-ROM.

Abry, Christian, Mohamed-Tahar Lallouache, and Marie-Agnes Cathiard. 1996. How can coarticulation models account for speech sensitivity to audio-visual desynchronization? In D. Stork and M. Hennecke (eds.), *Speechreading by Humans and Machines*, vol. 150, 247–256. Berlin: Springer-Verlag.

Ashby, W. Ross. 1956. *An Introduction to Cybernetics*. London: Chapman and Hall.

Barbosa, Adriano V. 2004. A study on the relations between audible and visible speech. PhD dissertation, Federal University of Minas Gerais, Belo Horizonte, Brazil.

Barbosa, Adriano V. and Eric Vatikiotis-Bateson. 2006. Video tracking of 2D face motion during speech. In F. Gebali and R. Ward (eds.), *IEEE Symposium of Signal Processing and Information Technology – ISSPIT 2006*, 1–6. Vancouver: IEEE.

Barbosa, Adriano V., Hani C. Yehia, and Eric Vatikiotis-Bateson. 2003. Modeling the relation between speech acoustics and 3D face motion. *Technical Report of the Institute of Electronics, Information, and Communication Engineers* 102(735): 13–18.

Barbosa, Adriano V., Hani C. Yehia, and Eric Vatikiotis-Bateson. 2008. Linguistically valid movement behavior measured non-invasively. In R. Goecke, P. Lucey, and S. Lucey (eds.), *Proceedings of the International Conference on Auditory-Visual Speech Processing 2008*, 173–177.

Barbosa, Adriano V., Rose-Marie Dechaine, Eric Vatikiotis-Bateson, and Hani C. Yehia. 2012. Quantifying time-varying coordination of multimodal speech signals using correlation map analysis. *Journal of the Acoustical Society of America* 131(3): 2162–2172.

Barbosa, Adriano V., Hani C. Yehia, Philip Rubin, and Eric Vatikiotis-Bateson. 2006. Relating the audible and visible components of speech. In H.C. Yehia, D. Demolin, and R. Laboissière (eds.), *Proceedings of the 7th International Seminar on Speech Production – ISSP 2006*, 119–126.

Bell-Berti, Fredericka and Katherine S. Harris. 1982. Temporal patterns of coarticulation: Lip rounding. *Journal of the Acoustical Society of America* 71: 449–454.

Bell-Berti, Fredericka, Thomas Baer, Katherine S. Harris, and Seiji Niimi. 1979. Coarticulatory effects of vowel quality on velar elevation. *Phonetica* 36: 187–193.

Browman, Catherine P. and Louis Goldstein. 1986. Towards an articulatory phonology. *Phonology Yearbook* 3: 219–252.

Bruce, Vicki and Andy W. Young. 1986. Understanding face recognition. *British Journal of Psychology* 77: 305–327.

Callan, Daniel E., Jeffrey A. Jones, Kevin Munhall, Akiko M. Callan, Christian Kroos, and Eric Vatikiotis-Bateson. 2004. Neural processes underlying perceptual enhancement by visual speech gestures. *NeuroReport* 14(17): 2213–2218.

Callan, Daniel E., Jeffrey A. Jones, Kevin G. Munhall, Christian Kroos, Akiko M. Callan, and Eric Vatikiotis-Bateson. 2002. Mirror neuron system activity and audiovisual speech perception. Paper presented at the Eighth International Conference on Functional Mapping of the Human Brain, Sendai, Japan.

Carpenter, Roger H.S. 1988. *Movement of the Eyes*, 2nd edn. London: Pion.

Carter, John N., Christine H. Shadle, and Colin J. Davies. 1996. On the use of structured light in speech research. Paper presented at the 1st ESCA Tutorial and Research Workshop on Speech Production Modeling: From Control Strategies to Acoustics and 4th Speech Production Seminar: Models and Data, Autrans, France.

Catford, John C. 1977. *Fundamental Problems in Phonetics*. Bloomington: Indiana University Press.

Cathiard, Marie-Agnes, Mohamed-Tahar Lallouache, and Christian Abry. 1996. Does movement on the lips mean movement in the mind? In D. Stork and M. Hennecke (eds.), *Speechreading by Humans and Machines*, vol. 150, 211–219. Berlin: Springer-Verlag.

Chomsky, Noam and Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper and Row.

dePaula, Hugo, Hani C. Yehia, Douglas Shiller, Gregoire Jozan, Kevin G. Munhall, and Eric Vatikiotis-Bateson. 2003. Linking production and perception through spatial and temporal filtering of visible speech information. Paper presented at the Fifth International Seminar on Speech Production, ISSP5, Macquarie University, Australia.

dePaula, Hugo, Hani C. Yehia, Douglas Shiller, Gregoire Jozan, Kevin G. Munhall, and Eric Vatikiotis-Bateson. 2006. Analysis of audiovisual speech intelligibility based on spatial and temporal filtering of visual speech information. In J. Harrington and M. Tabain (eds.), *Speech Production: Models, Phonetic Processes, and Techniques*, 135–147. London: Psychology Press.

Eigsti, Inge-Marie, Eric Vatikiotis-Bateson, Sumio Yano, and Kevin G. Munhall. 1995. Effects of listener expectation on eye movement behavior during audiovisual perception. *Journal of the Acoustical Society of America* 97: 3286.

Ekman, Paul. 1989. The argument and evidence about universals in facial expressions of emotion. In H. Wagner and A. Monstead (eds.), *Handbook of Social Psychophysiology*, 143–146. Chichester: John Wiley & Sons Ltd.

Ekman, Paul and Wallace V. Friesen. 1978. *Manual for the Facial Action Coding System*. Palo Alto, CA: Consulting Psychologists Press.

Ekman, Paul, Wallace V. Friesen, and Phoebe Ellsworth. 1972. *Emotion in the Human Face: Guidelines for Research and a Review of Findings*. New York: Pergamon Press.

Ewan, William G. 1979. Can intrinsic vowel F0 be explained by source/tract coupling? *Journal of the Acoustical Society of America* 66: 358–362.

Fant, Gunnar. 1960. *Acoustic Theory of Speech Production*. The Hague: Mouton.

Grant, Ken W. and Philip F. Seitz. 2000. The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America* 108: 1197–1208.

Hill, Harold and Eric Vatikiotis-Bateson. 2005. Using animations to investigate the perception of facial speech. Paper presented

at the ATR Symposium of Cross-Modal Processing of Faces and Voices, ATR Human Information Science Labs, Japan.
Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. *Neural Computation* 18: 1527–1554.
Horn, Berthold K.P. and Brian G. Schunk. 1981. Determining optical flow. *Artificial Intelligence* 17: 185–203.
Jakobson, Roman, Gunnar Fant, and Morris Halle. 1951/1963. *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*. Cambridge, MA: MIT Press.
Katz, William F. and Peter F. Assmann. 2001. Identification of children's and adults' vowels: Intrinsic fundamental frequency, fundamental frequency dynamics, and presence of voicing. *Journal of Phonetics* 29(1): 23–51.
Kelso, J.A. Scott, Betty Tuller, Eric Vatikiotis-Bateson, and Carol A. Fowler. 1984. Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance* 10: 812–832.
Kim, Midam, William S. Horton, and Ann R. Bradlow. 2011. Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Laboratory Phonology* 2(1): 125–156.
Kuratate, Takaaki, Eric Vatikiotis-Bateson, and Hana C. Yehia. 2005. Estimation and animation of faces using facial motion mapping and a 3D face database. In J.G. Clement and M.K. Marks (eds.), *Computer-Graphic Facial Reconstruction*, 325–346. Amsterdam: Academic Press.
Kuratate, Takaaki, Hana Yehia, and Eric Vatikiotis-Bateson. 1998. Kinematics-based synthesis of realistic talking faces. In D. Burnham, J. Robert-Ribes, and E. Vatikiotis-Bateson (eds.), *International Conference on Auditory-Visual Speech Processing 1998*, 185–190.
Massaro, Dominic W. and Michael M. Cohen. 1983. Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance* 9: 753–771.
McGurk, Harry and John MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264: 746–748.
Munhall, Kevin G. and Yoh'ichi Tohkura. 1998. Audiovisual gating and the time course of speech perception. *Journal of the Acoustical Society of America* 104: 530–539.
Munhall, Kevin G. and Eric Vatikiotis-Bateson. 1998. The moving face during speech communication. In R. Campbell, B. Dodd, and D. Burnham (eds.), *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech*, 123–139. Hove, UK: Psychology Press.
Munhall, Kevin G. and Eric Vatikiotis-Bateson. 2004. Spatial and temporal constraints on audiovisual speech perception. In G. Calvert, C. Spence, and B. Stein (eds.), *The Handbook of Multisensory Processes*, 177–188. Cambridge, MA: MIT Press.
Munhall, Kevin G., P. Gribble, L. Sacco, and M. Ward. 1996. Temporal constraints on the McGurk effect. *Perception and Psychophysics* 58(3): 351–362.
Munhall, Kevin G., Jeffrey A. Jones, Daniel E. Callan, Takaaki Kuratate, and Eric Vatikiotis-Bateson. 2004. Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science* 15(2): 133–137.
Munhall, Kevin G., Gregoire Jozan, Christian Kroos, and Eric Vatikiotis-Bateson. 2004. Spatial frequency requirements for audiovisual speech perception. *Perception and Psychophysics* 66(4): 574–583.
Pardo, Jennifer S. 2006. On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America* 119(4): 2382–2393.
Perkell, Joseph S., Marc Cohen, Mario A. Svirsky, Melanie Matthies, Inaki Garabieta, and Michel Jackson. 1992.

Electromagnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements. *Journal of the Acoustical Society of America* 92: 3078–3096.

Robert-Ribes, Jordi, Jean-Luc Schwartz, and Pierre Escudier. 1995. A comparison of models for fusion of the auditory and visual sensors in speech perception. *Artificial Intelligence Review* 9: 323–346.

Sumby, William H. and Irwin Pollack. 1954. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America* 26: 212–215.

Summerfield, Quentin. 1979. Use of visual information for phonetic perception. *Phonetica* 36: 314–331.

Summerfield, Quentin. 1987. Some preliminaries to a comprehensive account of audiovisual speech perception. In B. Dodd and R. Campbell (eds.), *Hearing by Eye: The Psychology of Lipreading*, 3–52. Hillsdale, NJ: Lawrence Erlbaum Associates.

Tekalp, A. Murat and Joern Ostermann. 2000. Face and 2-D mesh animation in MPEG-4. *Signal Processing: Image Communication* 15(4–5): 387–421.

Trager, George L. 1958. Paralanguage: A first approximation. *Studies in Linguistics* 13: 1–13.

Vatikiotis-Bateson, Eric and J.A. Scott Kelso. 1984. Remote and autogenic articulatory adaptation to jaw perturbations during speech. *Journal of the Acoustical Society of America* 75: S23–24.

Vatikiotis-Bateson, Eric and Kevin G. Munhall. 2012a. Empirical perceptual-motor linkage of multimodal speech. In G. Bailly, P. Perrier, and E. Vatikiotis-Bateson (eds.), *Advances in Auditory and Visual Speech Perception*, 346–367. Cambridge: Cambridge University Press.

Vatikiotis-Bateson, Eric and Kevin G. Munhall. 2012b. Time-varying coordination in multisensory speech processing. In B. Stein (ed.), *The New Handbook of Multisensory Processing*, 421–434. Cambridge, MA: MIT Press.

Vatikiotis-Bateson, Eric and David J. Ostry. 1995. An analysis of the dimensionality of jaw motion in speech. *Journal of Phonetics* 23: 101–117.

Vatikiotis-Bateson, Eric and Hani C. Yehia. 1996a. Physiological modeling of facial motion during speech. *Transactions Technical Committee Psychological Physiological Acoustics* H-96-65: 1–8.

Vatikiotis-Bateson, Eric and Hani C. Yehia. 1996b. Synthesizing audiovisual speech from physiological signals. Paper presented at the Acoustical Society of America and Acoustical Society of Japan Third Joint Meeting, 2–6 December, 1996, Honolulu, HI.

Vatikiotis-Bateson, Eric, Inge-Marie Eigsti, and Sumio Yano. 1994. Listener eye movement behavior during audiovisual perception. Paper presented at the International Conference on Spoken Language Processing (ICSLP-94), Yokohama, Japan.

Vatikiotis-Bateson, Eric, Inge-Marie Eigsti, Sumio Yano, and Kevin G. Munhall. 1998. Eye movement of perceivers during audiovisual speech perception. *Perception and Psychophysics* 60(6): 926–940.

Vilkman, Erkki, Olli Aaltonen, Ilkka Raimo, Paula Arajärvi, and Hanna Oksanen. 1989. Articulatory hyoid-laryngeal changes vs. cricothyroid muscle activity in the control of intrinsic F0 of vowels. *Journal of Phonetics* 17: 193–203.

Vitkovich, Melanie and Paul Barber. 1994. Effects of video frame rate on subjects' ability to shadow one of two competing verbal passages. *Journal of Speech, Language, and Hearing Research* 37: 1204–1210.

Woodward, Mary F. and Carroll G. Barber. 1960. Phoneme perception in lipreading. *Journal of Speech and Hearing Research* 3: 212–222.

Yehia, Hani C., Takaaki Kuratate, and Eric Vatikiotis-Bateson. 1999. Using speech acoustics to drive facial motion.

Paper presented at the 14th International Congress of Phonetic Sciences, San Francisco, CA.

Yehia, Hani C., Takaaki Kuratate, and Eric Vatikiotis-Bateson. 2002. Linking facial animation, head motion, and speech acoustics. *Journal of Phonetics* 30(3): 555–568.

Yehia, Hani C., Philip E. Rubin, and Eric Vatikiotis-Bateson. 1998. Quantitative association of vocal-tract and facial behavior. *Speech Communication* 26: 23–44.